DOCUMENT RESUME

ED 412 246 TM 027 497

AUTHOR Brooks, Gordon P.; Barcikowski, Robert S.
TITLE Precision Power Method for Selecting Regression Sample Sizes.
PUB DATE 1995-10-00
NOTE 46p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 1995).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Monte Carlo Methods; *Prediction; *Regression (Statistics); *Sample Size; Selection; Simulation
IDENTIFIERS Cross Validation; *Power (Statistics); *Precision (Mathematics)

ABSTRACT

When multiple regression is used to develop a prediction model, sample size must be large enough to ensure stable coefficients. If sample size is inadequate, the model may not predict well in future samples. Unfortunately, there are problems and contradictions among the various sample size methods in regression. For example, how does one reconcile differences between a 15:1 subject-to-variable ratio and a 30:1 rule. The purpose of this study was to validate a precision power method for determining sample sizes in regression. The method uses a cross-validity approach to selecting sample sizes so that models will predict as well as possible in future samples. The simple formula, which is an algebraic manipulation of a cross-validation formula, enables researchers to limit the expected shrinkage of R squared. Using a Monte Carlo simulation study, the precision power method was compared to eight other methods. It was the only method that provided consistently accurate and acceptable precision power rates. That is, when precision power was set a priori, actual precision power rates consistently fell within an acceptable interval around that given power rate. (Contains 3 tables and 78 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 412 246

Precision Power Method for Selecting Regression Sample Sizes

Gordon P. Brooks

Robert S. Barcikowski

Ohio University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

Gordon Brooks

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

TM027497

Paper presented at the annual meeting of the Mid-Western Educational Research Association, October 1995, Chicago, IL.

BEST COPY AVAILABLE

A New Sample Size Formula for Regression

ABSTRACT

When multiple regression is used to develop a prediction model, sample size must be large enough to ensure stable coefficients. If sample size is inadequate, the model may not predict well in future samples. Unfortunately, there are problems and contradictions among the various sample size methods in regression. For example, how does one reconcile differences between a 15:1 subject-to-variable ratio and a 30:1 rule?

The purpose of this study was to validate a precision power method for determining sample sizes in regression. The method uses a cross-validity approach to selecting sample sizes so that models will predict as well as possible in future samples. The simple formula, which is an algebraic manipulation of a cross-validation formula, enables researchers to limit the expected shrinkage of R^2.

Using a Monte Carlo simulation study, the precision power method was compared to eight other methods and was the only method which provided consistently accurate and acceptable precision power rates. That is, when precision power was set a priori, actual precision power rates consistently fell within an acceptable interval around that given power rate.

OBJECTIVES

Most researchers who use regression analysis to develop prediction equations are not only concerned with whether the multiple correlation coefficient or some particular predictor is

Paper presented at the annual meeting of the Mid-Western Educational Research Association, October 1995, Chicago, IL.

significant, but they are also especially concerned with the generalizability of the regression model developed. However, the process of maximizing the correlation between the observed and predicted criterion scores requires mathematical capitalization on chance; that is, the correlation obtained is a maximum only for the particular sample from which it was calculated. If the estimate of the population multiple correlation decreases too much in a second sample, the regression model has little value for prediction. Because of this possibility, researchers must ensure that their studies have adequate power so that results will generalize; the best way to ensure this power, and therefore stable regression weights, is to use a sufficiently large sample.

Despite encouragement from scholars, many researchers continue to ignore power in their studies (Cohen, 1992; Sedlmeier & Gigerenzer, 1989; Stevens, 1992b). This situation is compounded for multiple regression research even though several methods exist for choosing sample sizes for power. Unfortunately, as Olejnik noted in 1984 and was confirmed during the current research, many regression textbooks do not discuss the issue of sample size selection (e.g. Dunn & Clark, 1974; Kleinbaum, Kupper, & Muller, 1987; Montgomery & Peck, 1992; Weisberg, 1985) or simply provide a rule-of-thumb (e.g. Cooley & Lohnes, 1971; Harris, 1985; Kerlinger & Pedhazur, 1973; Tabachnick & Fidell, 1989), possibly because there are problems and contradictions among the various methods. These methods can be grouped loosely into three categories: rules-of-thumb, statistical power methods, and cross-validation methods. For example, how does one reconcile differences between a 15:1 subject-to-variable ratio and a 30:1 rule? Furthermore, the many rules-of-thumb lack any measure of effect size, which is generally recognized as a critical element in the determination of sample sizes. Cohen's (1988) methods are derived from a fixed model and statistical power approach to regression; however, a random

model and cross-validation approach, like Park and Dudycha's (1974), may be more appropriate in the social sciences, where a prediction function is often desired. This is because generalizability is the primary consideration for the development of a prediction model, whereas statistical power is the main concern when regression is used to test hypotheses about relationships between variables.

Therefore, the purpose of this paper is to reduce uncertainty about regression sample sizes. Through a Monte Carlo power study, a new and accessible method for calculating adequate sample sizes for multiple linear regression analyses will be validated. Because the new method, which will be called the precision power method, is developed primarily from a cross-validation approach, the next section begins from that perspective. However, because some aspects of the precision power method have been adapted from both the rules-of-thumb and the statistical power approach, a brief discussion of each will be included.

PERSPECTIVES

Cross-Validation and Shrinkage

Because the expected value of the sample multiple correlation (i.e. an average correlation over many samples) is an overestimate of the population multiple correlation, researchers have employed a number of methods to "shrink" R^2 and thereby provide better estimates of true population multiple correlations. Formula methods of shrinkage are typically preferred to empirical cross-validation (data-splitting) so that the entire sample may be used for model-building. Indeed, several common formula estimates have been shown superior to empirical cross-validation techniques (Cattin, 1980a; 1980b; Kennedy, 1988; Murphy, 1982; Schmitt, Coyle, & Rauschenberger, 1977).

Two types of formulas have been developed: shrinkage estimates and cross-validity estimates (see Table 1). Shrinkage formulas are used to estimate more accurately the squared population multiple correlation, ρ^2, also called the coefficient of determination. The multiple correlation, ρ, is the correlation between the criterion and the regression function if both are calculated in the population (Herzberg, 1969; Stevens, 1992a). For example, a researcher who calculates a sample $R^2=.3322$ with 121 subjects and 3 predictors might use an adjusted R^2 formula to conclude that, in the population, the multiple correlation between the criterion and the predictors is approximately $\rho=.5613$, since $R_a^2=.3151$.

Cross-validity formulas, which are based on estimates of the mean squared error of prediction, provide more accurate estimates of the squared population cross-validity coefficient, ρ_c^2. The values of R_c^2, the sample estimates of cross-validity, will vary from sample to sample; however, the expected value of R_c^2 (that is, the average over many samples) approximates ρ_c^2. This cross-validity coefficient can be thought of as the squared correlation between the actual population criterion values and the scores predicted by the sample regression equation when applied to the population or to another sample (Kennedy, 1988; Schmitt et al., 1977). For example, a researcher who calculates a sample $R^2=.3322$ with 121 subjects and 3 predictors might use a cross-validity formula to calculate the sample cross-validity coefficient as $R_c^2=.2916$. This cross-validity coefficient implies that the researcher would explain 29%, not 33%, of the variance of the criterion when applying the sample regression function to future samples.

The most common estimate of shrinkage reported in the literature (and in statistical packages) is an adjusted R^2 that is attributed most frequently to Wherry (1931). However, when researchers are interested in developing a regression model to predict for future subjects, they

should report both R_a^2 (for descriptive purposes) and R_c^2, which indicates how well their sample equation may predict in subsequent samples (Cattin, 1980b; Huberty & Mourad, 1980). Indeed, Uhl and Eisenberg (1970) found that a cross-validity estimate (which they attribute to Lord, 1950) was consistently more accurate than Wherry's shrinkage formula in this regard. Some of the more familiar cross-validity formulas are those by Stein (1960), Darlington (1968), Lord (1950), Nicholson (1960), and Browne (1975).

Cross-Validation Approach to Sample Sizes

Park and Dudycha (1974) took a cross-validation approach to calculating sample sizes. They noted that such a cross-validation approach is applicable to both the random and the fixed models of regression; however, because the fixed model poses no practical problems, they emphasized the random model. In the random model, both the predictors and the criterion are sampled together from a joint multivariate distribution. The fixed model, on the other hand, assumes that the researcher is able to select or control the values of the independent variables before measuring subjects on the random dependent variable. The random model is usually more appropriate to social scientists because they typically measure subjects on predictors and the criterion simultaneously and therefore are not able to fix the values for the independent variables (Brogden, 1972; Cattin, 1980b; Claudy, 1972; Drasgow, Dorans, & Tucker, 1979; Herzberg, 1969; Park & Dudycha, 1974; Stevens, 1986, 1992a). For a more complete discussion of the random and fixed models, the reader is referred to Afifi and Clark (1990), Brogden (1972), Dunn and Clark (1974), Johnson and Leone (1977), and Sampson (1974).

Park and Dudycha derived the following sample size formula: $N \geq [(1-\rho^2)\delta_1^2/\rho^2]+p+2$, where δ_1^2 is the noncentrality parameter for the t-distribution. Researchers determine the

probability with which they want to approximate ρ within some chosen error tolerance. The formula for this probability is: $P(\rho-\rho_c \leq \epsilon)=\gamma$. The researcher chooses (a) an assumed ρ^2 as the effect size, (b) the absolute error willing to be tolerated, ϵ, and (c) the probability of being within that error bound, γ. The tables provided by Park and Dudycha (most of which were reprinted in Stevens, 1986, 1992a) can then be consulted with these values. Unfortunately, their tables are limited to only a few possible combinations of sample size, squared correlation, and epsilon. Also unfortunately, their math is too complex for most researchers to derive the information they would need for the cases not tabulated. Additionally, there is no clear rationale for how to determine the best choice of either epsilon or the probability to use when consulting the tables (although Stevens, 1992a, implied through examples that .05 and .90, respectively, are acceptable values).

Rules of Thumb for Selecting Sample Sizes

The most extensive literature regarding sample sizes in regression analysis is in the area of experiential rules. Many scholars have suggested rules-of-thumb for choosing sample sizes that they claim will provide reliable estimates of the population regression coefficients. That is, with a large enough ratio of subjects to predictors, the estimated regression coefficients will be reliable and will closely reflect the true population parameters since shrinkage will be slight (Miller & Kunce, 1973; Pedhazur & Schmelkin, 1991; Tabachnick & Fidell, 1989). This is true because as the number of subjects increases relative to the number of predictors, both R^2 and ρ_c^2 converge toward ρ^2, and therefore the amount of shrinkage decreases (Cattin, 1980a).

Rules-of-thumb typically take the form of a subject-to-predictor (N/p) ratio. Table 2 shows that statisticians have recommended using as small a ratio as 10 subjects to each predictor

and as large a ratio as 40:1. For example, Stevens (1986) recommended a 15:1 subject-to-variable ratio, which he based primarily on an analysis of Park and Dudycha's (1974) tables. Harris (1985) noted, however, that ratio rules-of-thumb clearly break down for small numbers of predictors. Some scholars have suggested that a minimum of 100, or even 200, subjects is necessary regardless of the number of predictors (e.g. Kerlinger & Pedhazur, 1973). Indeed, Green (1991) found that a combination formula such as $N>50+8p$ was much better than subject-to-variable ratios alone. Additionally, Sawyer (1982) developed a formula based on limiting the inflation of mean squared error. Sawyer's formula, $N \geq [(2k^2-1)+k^2p]/(k^2-1)$, easily simplifies into a combination rule once the inflation factor, k, is chosen. Finally, perhaps the most widely used rule-of-thumb was described by Olejnik (1984): "use as many subjects as you can get and you can afford" (p. 40).

The most profound problem with many rules-of-thumb advanced by regression scholars is that they lack any measure of effect size. Indeed, even Sawyer's inflation factor is not an effect size. It is generally recognized that an estimated effect size must precede the determination of appropriate sample size. Effect size enables a researcher to determine in advance not only what will be necessary for statistical significance, but also what is required for practical significance (Hinkle & Oliver, 1983). The next section includes a more complete discussion of effect size and its importance in statistical power analysis.

Statistical Power Approach

"The power of a statistical test is the probability that it will yield statistically significant results" (Cohen, 1988, p. 1). That is, statistical power is the probability of rejecting the null hypothesis when the null hypothesis is indeed false. Statistical power analysis requires the

consideration of at least four parameters: level of significance, power, effect size, and sample size. These four parameters are related such that when any three are fixed, the fourth is mathematically determined (Cohen, 1992). Therefore, it becomes obvious that it is necessary to consider power, alpha, and effect size when attempting to determine a proper sample size. This is a fixed model approach to regression, however, and is most useful when researchers use regression as a means to explain the variance of a phenomenon in lieu of analysis of variance. It is useful, though, to discuss effect size regardless of the approach to regression that is taken.

In any statistical analysis, there are three strategies for choosing an appropriate effect size: (a) use effect sizes found in previous studies, (b) decide on some minimum effect that will be practically significant, or (c) use conventional small, medium, and large effects (Cohen & Cohen, 1983). Cohen (1988) defined effect size in fixed model multiple regression as a function of the squared multiple correlation, specifically $f^2=R^2/(1-R^2)$. Since R^2 can be used in the formulas directly, Cohen also defined effect sizes in terms of R^2 such that small effect $R^2=.02$, medium effect $R^2=.13$, and large effect $R^2=.26$. Cohen's (1988) sample size is calculated as $N=\lambda(1-R^2)/R^2$, where λ is the noncentrality parameter required for the noncentral F-distribution. Cohen's (1988) tables provide the λ needed for the sample size formula.

For prediction studies, the fundamental problem with Cohen's (1988) method, and Green's (1991) formula based on Cohen's method, is that it is designed for use from a fixed model, statistical power approach. And although Gatsonis and Sampson (1989) use the random model approach, their method is also based on a statistical power approach to sample size determination. Unfortunately, statistical power to reject a null hypothesis of zero multiple correlation does not inform us how well a model may predict in other samples. That is, adequate sample sizes for

statistical power tell us nothing about the number of subjects needed to obtain stable, meaningful regression weights (Cascio, Valenzi, & Silbey, 1978). Therefore, selecting a sample size based on statistical power tests may be useful in selecting predictors to include in a final model, but it will not ensure adequate sample size to allow a regression equation to generalize to other samples from the given population.

Preliminary Study: The Rozeboom-based Method

Because the methods described above (a) provide contradictory sample size recommendations, (b) either oversimplify the issue or are too mathematically complex for many researchers to use, and (c) are not all based on the random model, a new sample size selection method for multiple regression was developed. The cross-validity formula in Table 1 developed by Rozeboom (1978) was adapted to form the sample size formula:

$$N \geq [p\,(2 - 2R^2 + \epsilon)] / \epsilon, \qquad (1)$$

where p is the number of predictors, R^2 is the expected sample value, and ϵ is the acceptable absolute amount of shrinkage, such that $\epsilon = R^2 - R_c^2$.

A Monte Carlo power study was conducted to determine the efficacy of the new Rozeboom-based method as compared to several existing methods of selecting regression sample sizes. A more complete discussion of this preliminary study was presented in Brooks and Barcikowski (1994). The Rozeboom-based method and Park and Dudycha's (1974) method consistently produced the highest rates of predictive power. However, closer inspection of the results indicates two undesirable characteristics of all the methods examined. First, although the relative rankings of the methods remained fairly consistent across predictors, their absolute power rates did not. For example, with expected $R^2 = .25$ and $\rho^2 \approx .25$, the Rozeboom-based method

provided a predictive power rate of .71 with two predictors, .75 with three predictors, .76 with four predictors, and .80 with both eight and 15 predictors. Park and Dudycha's method and both the 15:1 and 30:1 rules exhibited similar results. Interestingly, Cohen's method, Gatsonis and Sampson's method, and the combination formula N=50+8p showed the opposite trend; that is, power decreased as the number of predictors increased.

Second, the predictive power rates of all but Cohen's method decreased as the expected R^2 decreased when expected $R^2 \approx \rho^2$ (Cohen's method provided its highest power rates when expected R^2 was smallest). For example, with three predictors, the Rozeboom-based method provided PP=.88 when $\rho^2 \approx .50$ and $R^2 = .50$, PP=.75 when $\rho^2 \approx .25$ and $R^2 = .25$, and PP=.64 when $\rho^2 \approx .10$ and $R^2 = .10$. Therefore, appropriate sample sizes were often too large when expected R^2 was larger, but were underestimated for smaller expected R^2 values.

The theory behind the method was that the researcher, knowing shrinkage was likely to occur, could set a limit as to the amount of shrinkage that would result. However, Rozeboom's formula overestimates both of the most widely accepted formulas, Darlington-Stein (Darlington, 1968; Stein, 1960) and Lord-Nicholson (Lord, 1950; Nicholson, 1960). Therefore, the shrinkage that occurred when using the Darlington-Stein formula was greater than the Rozeboom-based formula suggested. For example, in the case of two predictors with expected $R^2 = .25$ and $\rho^2 \approx .25$, cross-validity shrinkage (what Stevens would call "loss in predictive power") was .05 as expected when calculated with the Rozeboom formula, but was .06 when using Darlington-Stein. Since the definition of power used the Darlington-Stein estimate, the power rates were not as accurate as expected. In the example, predictive power was found to be .75 but was expected to be .80.

The New Precision Power Method

The preliminary study confirmed that the Rozeboom-based method provided consistently higher power rates than the other methods examined. However, informed choice of sample size requires that researchers be able to set power a priori, not simply use the method which provides the highest power. Closer scrutiny has shown that all existing methods, including the Rozeboom-based method, share the inconsistencies described above across number of predictors and across expected R^2 values. As a result, none of the methods provides a reliable means to set power in advance of the research. Therefore, the purpose of this paper was to extend the analysis to determine if any method can provide consistently *accurate* power rates, not simply the highest rates. A new sample size formula was developed based on Lord (as cited in Uhl & Eisenberg, 1970). (It should be noted that the Lord formula cited by Uhl and Eisenberg (1970) differs from the most common interpretation of Lord's 1950 paper, which is represented in Table 1). The Lord formula was chosen primarily because it provides cross-validity estimates that typically underestimate the Darlington-Stein values rather than overestimate them, as does the Rozeboom formula.

Like Rozeboom's (1978) cross-validity formula, Lord's "relatively unknown formula" (Uhl & Eisenberg, 1970, p. 489) is linear in all parameters, which makes it ideal for algebraic manipulation:

$$R_c^2 = 1 - [(N+p+1)(1-R^2) / (N-p-1)], \quad (2)$$

where N is sample size, p is the number of predictors, and R^2 is the actual sample value. Uhl and Eisenberg (1970) found this formula to give accurate estimates of "cross-sample" shrinkage,

regardless of sample size and number of predictors. Algebraic manipulation and simplification of the formula to solve for sample size yields:

$$N = [(p+1)(2-2R^2+\epsilon)] / \epsilon, \quad (3)$$

where p is the number of predictors, R^2 is the expected sample value, and ϵ is an acceptable amount of shrinkage (i.e. $\epsilon = R^2 - R_c^2$). This value of ϵ allows researchers to decide how closely to estimate ρ_c^2 from expected R^2, as it did in the Rozeboom-based formula.

This new precision power formula (3) includes the same effect size parameter, in the form of expected R^2, as the Rozeboom-based method. The precision power formula also has the same capacity for simplification as the Rozeboom-based formula: either an absolute amount of acceptable shrinkage (e.g. $\epsilon = .05$) or a proportional decrease (e.g. $\epsilon = .2R^2$, which represents shrinkage of 20%). For example, if a researcher wanted an estimate of ρ_c^2 not less than 80% of the sample R^2 value, the formula (3) can be reformulated using $\epsilon = R^2 - .8R^2 = .2R^2$, such that

$$N \geq [(p + 1)(2 - 1.8R^2)] / 0.2R^2; \quad (4)$$

or if the researcher wanted a ρ_c^2 estimate not less than 75% of the sample R^2 value, the formula would be reformulated such that $\epsilon = .25R^2$:

$$N \geq [(p + 1)(2 - 1.75R^2)] / 0.25R^2. \quad (5)$$

If the researcher did not want the sample R^2 to decrease by more than .05 no matter what the expected value of R^2, formula (3) simplifies to

$$N \geq 20 (p + 1)(2.05 - 2R^2); \quad (6)$$

or if the researcher did not want the sample R^2 to decrease by more than .03, then

$$N > 33 (p + 1)(2.03 - 2R^2). \quad (7)$$

Because the Lord-based precision power sample size method is based on a cross-validation formula which more closely approximates the Darlington-Stein formula than does the Rozeboom cross-validity formula, it was expected that precision power rates would be more consistently accurate. However, there is no way to compare this method to current methods mathematically. Therefore, a Monte Carlo power study was performed to determine the efficacy of the new method as compared to several existing methods, especially the Rozeboom-based method. The next section describes this study in detail.

METHODS AND DATA

Ideally, a mathematical proof would be provided that would compare directly the efficacies of existing sample size methods and the new precision power method offered in this paper (Halperin, 1976; Harwell, 1990). However, the several sample size selection methods compared here are based on different probability distributions, making direct comparison problematic. For example, Park and Dudycha (1974) base their work on the probability density function of ρ_c^2, Cohen (1988) bases his material on the noncentral χ^2 distribution, Gatsonis and Sampson (1989) base their method on the distribution of R_{xy}, and rules-of-thumb are not based on probability distributions at all.

Fortunately, meaningful comparisons among the power rates of these methods can be accomplished through a Monte Carlo study. Monte Carlo methods use computer assisted simulations to provide evidence for problems that cannot be solved mathematically. In Monte Carlo statistical power studies, random samples are generated and used in a series of simulated experiments in order to calculate empirical power rates. That is, many random samples are generated such that the null hypothesis is known to be false (e.g. the multiple correlation is non-

null) and then the actual number of tests that are correctly rejected are counted. After all samples are completed, a proportion is calculated that represents the actual statistical power rate. This methodology is adapted here to apply to the precision power being studied.

Definition of Precision Power

While several scholars have used the term *predictive power* (e.g. Cascio et al., 1978; Kennedy, 1988; Stevens, 1986, 1992a), only Cattin (1980a) has provided a formal definition. Cattin (1980a) noted that the two common measures of predictive power are the mean squared error of prediction and the cross-validated multiple correlation. However, Cattin was discussing predictive power in regard to the comparison and selection of competing regression models. Stevens (1986, 1992a), who discussed predictive power as an aspect of model validation, used the term to mean how well a derived regression equation will predict in other samples from the same population. Therefore, a "loss in predictive power" to Stevens is simply the size of the decrease in the sample R^2 when an appropriate shrinkage or cross-validity formula is applied.

Although both Cattin's and Steven's definitions of predictive power could be applied to the current circumstances in some fashion, neither would provide any sense of the magnitude of error as compared to the original R^2 value. For example, a loss in predictive power (as Stevens defines it) of .20 suggests drastically different results if the sample R^2 is .50 than if the sample R^2 is .25. Because they desire a regression model that predicts well in subsequent samples, researchers hope to limit shrinkage as much as possible relative to the sample R^2 value they attained. Therefore, a concept is required that provides more information about the magnitude of shrinkage relative to sample values.

Therefore the term *precision power* has been defined for this study to indicate how well a regression function is expected to perform if applied to future samples. The term is adapted from Darlington (1990), who used the phrase "precision of estimates" to oppose the "power of hypothesis tests" during his introduction to a chapter on choosing sample sizes (p. 379). Precision power is defined more precisely as R_c^2/R^2, which can be inferred and adapted from an example used by Stevens (1992a, p. 100). With a larger sample, this fraction would be larger because less shrinkage occurs with larger samples, all else remaining constant. Using Stevens' example, a 61.8% shrinkage from sample $R^2=.50$ to $R_c^2=.191$ occurs with a sample size of 50; when the sample is increased to 150, there is only a 15.8% shrinkage from $R^2=.50$ to $R_c^2=.421$. The precision power, as defined in this paper, in the first case would be .191/.50=.382, and precision power in the second case is .421/.50=.842.

The definition of precision power,

$$PP = R_c^2/R^2, \tag{8}$$

can be used a priori with the newly developed sample size method by algebraically manipulating the formula (8) and by setting R^2 equal to the expected value of R^2 in the population. By adding and subtracting one from the fraction,

$$PP = 1 - (R^2/R^2) + (R_c^2/R^2). \tag{9}$$

Combining the fractions provides

$$PP = 1 - (R^2 - R_c^2)/R^2. \tag{10}$$

The fraction which now remains, $(R^2-R_c^2)/R^2$, can be interpreted as the proportional decrease, or proportional shrinkage (PS), in the squared multiple correlation after an appropriate cross-validity estimate is made. Therefore, 1-PS provides an estimate of the precision power of the regression

equation. The derivation of sample size formula (3) substitutes for the quantity R^2-R_c^2 with ϵ, which represents the shrinkage tolerance as either absolute (e.g. ϵ=.05) or relative (e.g. ϵ=.2R^2). The relationship between the two formulas becomes obvious: the numerator in formula (10) also represents the shrinkage tolerance level, using a sample value for the cross-validity coefficient rather than an estimated population value. Therefore, formula (10) can be rewritten as

$$PP = 1 - \epsilon/R^2 \qquad (11)$$

and therefore

$$\epsilon = R^2 - (PP * R^2) \qquad (12)$$

to use similar variables as formula (3). For example then, if researchers wanted the R_c^2 after shrinkage to be no less than 80% of the expected sample R^2 of .50 with four predictors, they would set PP=.80, and therefore choose ϵ=.10 to use in sample size formula (3). Plugging the values into formula (3) provides a sample size of N=[5(2-2(.50)+.10)]/.10=55. Thus, 55 subjects should provide a large enough sample so that R_c^2>.40, which is 80% of the assumed ρ^2=.50.

Precision power thus describes how well the regression equation will predict in other samples relative to its ability to predict in the derivation sample. For example, a predetermined acceptable PS level of .20 provides precision power of .80. To carry the example out fully, precision power of .80 indicates that the sample was large enough to allow the sample R^2 to shrink by only 20%. To provide a numerical example, if sample R^2=.50 and R_c^2=.40, the sample value has shrunk only by 20%; whereas, a smaller sample size may cause R^2=.50 to shrink to R_c^2=.30, which is a 40% decrease and leads to a precision power of only .60.

Because the term *power* has special meaning in the research literature, a word of warning may be prudent at this time. Precision power as defined here, 1-PS, looks similar in form to the

theoretical definition of statistical power, 1-ß, where ß is the probability of a Type II error. However, PS is not the probability of error but the tolerance level for error, or more precisely, shrinkage. Furthermore, the term statistical power is used in reference to a test of an hypothesis; the term precision power, on the other hand, applies not to a statistical test, but to an evaluation of the generalizability of a regression equation.

The Stein (1960) cross-validity formula (sometimes attributed to Darlington, 1968 and Herzberg, 1969) was used for R_c^2 because it is has been recommended by many scholars who have investigated cross-validation techniques from a random model perspective (e.g. Claudy, 1978; Huberty & Mourad, 1980; Kennedy, 1988; Schmitt et al., 1977; Stevens, 1986, 1992a). It should be noted that the authors are aware that the Stein formula is not uniformly regarded as the best cross-validation formula (e.g. Cattin, 1980a; Darlington, 1990; Drasgow et al., 1979; Rozeboom, 1978). Statistical power was calculated as the proportion of total number of correct rejections to the total tests performed for each testing situation.

Research Design

A Monte Carlo analysis of the precision power rates of several regression sample size methods was performed. Because a variety of factors may influence precision power, several testing situations were considered. Four factors were manipulated and fully crossed for the present study. First, four effect sizes were used which represented the expected R^2: .10, .25, .50, and .75. The .10 and .25 values were chosen because they are found in Park and Dudycha's (1974) tables and because they are very close to Cohen's (1988) medium and large effect sizes of .13 and .26, respectively. The .50 value was chosen because Stevens (1992a) recommends it as "a reasonable guess for social science research" (p. 125). The .75 value was chosen to include a

value considered very large for comparison and completeness. Second, data were generated for five sets of predictors: 2, 3, 4, 8, 15. Again, these numbers were chosen for ready comparison with tables provided by both Park and Dudycha (1974) and Gatsonis and Sampson (1989) and because they provide a wide range of predictor numbers. Third, six separate values for the true population ρ^2 were used: .00, .05, .10, .25, .50, and .75. For ρ^2=.00, the appropriate identity matrix was used; for the other values, correlation matrices were created with R^2 values within ±.005 of these values using a procedure described in the Data Source section below. The population correlation values of .10, .25, .50, and .75, were chosen because they are the effect sizes chosen above. The .05 value was chosen to provide a population value lower than the .10 expected R^2 value and .00 was chosen to verify Type I error rates.

Finally, nine sample size selection methods were compared: the new Lord-based precision power method, the Rozeboom-based method, a method based on Sawyer (1982), Park and Dudycha (1974), Cohen (1988), Gatsonis and Sampson (1989), the 30:1 subject-to-variable ratio from Pedhazur and Schmelkin (1991), the $N \geq 50+8p$ formula from Green (1991), and the 15:1 ratio from Stevens (1992a). For both the precision power method and also the Rozeboom-based method, the value for ϵ was set both absolutely and proportionally (.05 and $.2R^2$, respectively). For Park and Dudycha's method, $P(\rho-\rho_c \leq \epsilon)=.90$ and $P(\rho-\rho_c \leq \epsilon)=.95$ were both included. Therefore, a total of 12 different methods were included in the analysis.

Sawyer (1982) considers that the inflation factor, k, should be set as a constant. If k is set as a constant, though, Sawyer's method simplifies to a rule-of-thumb. Therefore, for this study, Sawyer's method was adapted in an attempt to provide the method with an effect size. Sawyer's formula works such that as the inflation factor decreases, so does MSE, which also decreases

shrinkage. The value of k was set such that $k=1+\epsilon$, where $\epsilon=R^2-R_c^2$ or more specifically, $\epsilon=R^2-.8R^2$. This is the same calculation for ϵ as described for the precision power method in a previous section.

Monte Carlo Procedures

Turbo Pascal 6.0 code was written to calculate the sample sizes for the new method, the ratio methods, the combination method from Green (1991), and Cohen's (1988) method (after looking up the stored tabulated lambda values). The relevant sample size tables from both Park and Dudycha (1974) and Gatsonis and Sampson (1989) were stored as data for access by the computer program, as were the appropriate tables for Cohen's lambda values. Where the precision power method and the Rozeboom-based method were set absolutely, the value of ϵ was set to .03 for expected $R^2 \leq .10$ and $\epsilon=.05$ for expected $R^2>.10$; this is also the way Park and Dudycha's tables were handled. Both Cohen's and Gatsonis and Sampson's tables were entered using power=.90. It should be noted that for the case of expected $R^2=.50$, the tabulated sample size for $\rho=.70$ from Gatsonis and Sampson was used; for the case of expected $R^2=.10$, the $\rho=.30$ value was chosen; and for expected $R^2=.75$, the table values for $\rho=.85$ were used. Because in each of these cases the ρ value used was less than the square root of the expected R^2, the sample sizes chosen for the Gatsonis and Sampson method were slightly larger than exact values would have provided. The 12 methods do provide a variety of suggested sample sizes, sometimes drastically different (see Table 3).

A Turbo Pascal 6.0 program was written that generated and tested 15,614 samples for each of the above 1440 conditions. The number of iterations was chosen based on the work of Robey and Barcikowski (1992), who suggested that 15,614 iterations are needed for nominal

α=.05, nominal power of .80, Type I error rate for the two-tailed proportions test of .05, and the fairly stringent magnitude of departure $\alpha \pm .1\alpha$. For each iteration's sample, the program performed a standard regression analysis (all predictors entered simultaneously), calculated the necessary statistics and probabilities, tested the null hypothesis of zero correlation at a .05 significance level, and calculated the shrinkage/cross-validity estimates and precision power rates needed for the study.

Because the null hypothesis (H_0: ρ=0) was known to be false in each sample where ρ^2>.00, each rejection at a .05 significance level qualified as a correct rejection and was recorded as such. For each of these conditions, then, empirical statistical power rates were calculated simply as the proportion of the 15,614 tests that were correctly rejected. Also for each condition, average shrinkage and average cross-validity were calculated. Additionally, precision power for each condition was calculated as the average of the ratios of the Stein cross-validity coefficient to the sample R^2. For the occasions when the Stein estimate was negative, precision power was set to zero, which is its theoretical minimum. Finally, these summary data were compared to determine how well the sample size methods performed both absolutely and relatively. Simulated samples were chosen randomly to test program function by comparison with results provided by SPSS/PC+ version 5.0.1. Additionally, Type I error rates were found under the condition where population ρ^2=.00, also to validate program functioning. In only one of the 240 null-case samples tested did the empirical error rate fall outside of Bradley's (1978) "fairly stringent criterion" $\alpha \pm .1\alpha$, or $.045 \leq \alpha \leq .055$. Further, precision power rates were larger than .05 in only 2 of the 240 null cases.

The program was run as a MS-DOS 6.2 application under Windows 3.1 on a computer equipped with an updated Intel Pentium-100 processor, which has a built-in numeric processor. Extended precision floating point variables, providing a range of values from 3.4×10^{-4932} to 1.1×10^{4932} with 19 to 20 significant digits, were used.

DATA SOURCE

Because this research focused on power for the random model of regression, data were generated to follow a joint multivariate normal distribution. The first step was to create population correlation matrices that met the criteria required by this study, namely, appropriate numbers of variables and appropriate ρ^2 values. These correlation matrices were then used to generate multivariate normal data following a Cholesky decomposition procedure recommended by several scholars (Chambers, 1977; Collier, Baker, Mandeville, & Hayes, 1967; International Mathematical and Statistical Library, 1985; Karian & Dudewicz, 1991; Kennedy & Gentle, 1980; Keselman, Keselman, & Shaffer, 1991; Morgan, 1984; Ripley, 1987; Rubinstein, 1981).

For each range of ρ^2 and number of predictors (25 total conditions), a correlation matrix was created using the following procedure. Uniform random numbers between 0.0 and 1.0 were generated using a subtractive method algorithm suggested by Knuth (1981) and coded in Pascal by Press, Flannery, Teukolsky, and Vetterling (1989). These values were entered as possible correlations into a matrix and the squared multiple correlation, R^2, was calculated. If the R^2 value fell in the required range, the matrix was then tested to determine whether it was positive definite. Press, Teukolsky, Vetterling, and Flannery (1992) suggested that the Cholesky decomposition is an efficient method for performing this test -- if the decomposition fails, the matrix is not positive definite. The algorithm for the Cholesky decomposition used in this study was adapted from Nash

(1990). This procedure was repeated until the necessary 25 matrices were created. These correlation matrices were then used to generate the random samples as described below. It is worthwhile to note that with given values of R^2, sample size, and numbers of predictors, the distribution of the squared cross-validity coefficient does not depend on the particular form of the population covariance, or in this case correlation, matrix (Drasgow et al., 1979).

The Cholesky decomposition of a matrix produces a lower triangular matrix, L, such that $LL^T=\Sigma$, where Σ is a symmetric, positive definite matrix such as a covariance or correlation matrix. This lower triangular matrix, L, can be used to create multivariate pseudorandom normal variates through the equation

$$Z_{ij} = \mu_j + XL^T \qquad (13)$$

where Z_{ij} is the multivariate normal data matrix, μ_j is the mean vector, and X contains vectors of independent, standard normal variates. When $\mu_j=0$, the multivariate pseudorandom data is distributed with mean vector zero and covariance matrix Σ. Independent pseudorandom normal vectors, X_j, with means, zero, and variances, unity, were generated using an implementation of the Box and Muller (1958) transformation adapted from Press, Flannery, Teukolsky, and Vetterling (1989). The Box and Muller algorithm converts randomly generated pairs of numbers from a uniform distribution into random normal deviates.

RESULTS

The primary concern of this study was whether one or more of the methods examined provides consistently accurate precision power rates. That is, does any method of selecting sample sizes for regression recommend sample sizes that guarantee a certain level of precision

power regardless of the number of predictors and the value of expected R^2? The results discussed below also include confirmation of several results from the preliminary study.

Accuracy of the Methods

In order to answer the question of which method provides the most consistently accurate precision power (PP) rates, results of the 12 methods were compared using an adaptation of Robey and Barcikowski's (1992) intermediate criterion for robustness. Specifically, the accuracy of the level of proportional shrinkage (PS) was tested using the criterion PS±¼PS. For example, when precision power is expected to be .80, proportional shrinkage is expected to be .20. The interval for acceptable accuracy is therefore .20±.05, or .15≤PS≤.25; in terms of precision power, the acceptable interval is .75≤PP≤.85. (It should be noted that the Stein cross-validity estimates were verified to fall within the expected ranges based on the precision power rates found in the study.)

The preliminary study (Brooks & Barcikowski, 1994) showed that when researchers choose an expected R^2 that overestimates ρ^2 (either explicitly by choice of an inflated effect size or implicitly by use of an inappropriate rule-of-thumb), power rates are unacceptably low. Similarly, when researchers choose an expected R^2 which is much lower than the population ρ^2, power rates are unnecessarily high (more subjects than necessary are recommended). Results from the current study corroborate these findings, thereby reinforcing the need for thoughtful choice of effect size in regression research. For example, with four predictors and a population ρ^2=.25, and using sample sizes recommended by the precision power method results in the following scenarios: (a) with 22 subjects for expected R^2=.75, average PP=.18 and the average Stein cross-validity estimate shrinks from R^2=.38 to R_c^2=.0008; (b) with N=55 for expected

R^2=.50, PP=.50 and Stein shrinks from R^2=.30 to R_c^2=.17; (c) with N=155 for expected R^2=.25, PP=.82 and Stein shrinks from R^2=.27 to R_c^2=.23; and (d) with 455 subjects recommended for expected R^2=.10, PP=.94 and the Stein cross-validity estimate shrinks from R^2=.26 to R_c^2=.24.

Consequently, the only cases of real interest for a discussion of accuracy are those cases where the researcher has made a reasonable estimate of ρ^2. In all 20 conditions where expected $R^2=\rho^2$, both the proportional precision power method ($\epsilon=.2R^2$) and the absolute precision power method ($\epsilon=.05$) provided PP rates between within the interval PS±¼PS. The expected PP rates for the proportional precision power method remain at .80 regardless of the other conditions; therefore the acceptable interval for PP rates was $.75 \leq PP \leq .85$. Note, however, that PP rates for the absolute method change as expected R^2 changes (e.g. if $\epsilon=.05$ and expected R^2=.50, expected PP=.45/.50=.90, but PP=.80 if $\epsilon=.05$ and expected R^2=.25). Therefore, when expected R^2 was .75, the acceptable PP rates fell within the range $.917 \leq PP \leq .95$, where PP was expected to be .933.

The proportional Rozeboom-based method provided PP rates within the range $.75 \leq PP \leq .85$ in 13 of the 20 cases where $R^2=\rho^2$. The absolute method provided PP rates within ±¼PS of their expected rates in 15 of the 20 cases. Park and Dudycha's method at a probability of .95 provided PP rates within ±¼PS in 9 of 20 cases; at a probability of .90, Park and Dudycha's method did not provide any accurate PP rates. Sawyer's method, as adapted for this study, provided accurate rates in the interval $.75 \leq PP \leq .85$ only in the five cases for expected R^2=.75. Both Cohen's method and that of Gatsonis and Sampson provided PP rates much below what was expected. Finally, the remaining methods, based on rules-of-thumb, did not provide consistent

empirical PP rates over any set of conditions (which would have been required since none of these methods provides a means to choose a PS value).

Supplemental Analysis

After determining which methods provide accurate and consistent precision power rates at the levels tested above, a supplemental analysis was run to examine three of these methods at several different levels of precision power and expected $R^2=\rho^2$. Specifically, the precision power method, the Rozeboom-based method, and a new adaptation of Sawyer's (1982) method were analyzed using the same method described above for creating data. Because the adaptation of Sawyer's (1982) method used in the primary analysis was not as fruitful as desired, a different choice was made for the inflation factor, k, in the supplemental analysis: the inflation factor was set as $k=1+.1R^2$ for the supplemental analysis. While this inflation factor does not provide a means to set acceptable precision power a priori as did the original inflation factor, it will generally provide larger sample sizes than the original *k* and therefore provide higher empirical precision power rates.

In this supplemental analysis, only the cases where the researcher estimates an effect size for ρ^2 correctly were included. Additionally, only 2660 iterations were performed, which meets Robey and Barcıkowski's (1992) recommendation for an intermediate criterion of robustness ($\alpha \pm \frac{1}{4}\alpha$) when nominal $\alpha = .05$, nominal power is .80, and the Type I error rate for the two-tailed proportions test is .05. There were a total of 288 conditions examined such that (a) a priori precision power rates were varied from .60 to .90 by .10, (b) the number of predictors were varied from 1 to 9, inclusively, and (c) the expected $R^2=\rho^2$ values were varied from .20 to .90 by .10.

The supplemental data were analyzed in the same manner as the primary data. That is, the accuracy of the level of proportional shrinkage (PS) was tested using the criterion PS±¼PS. The precision power method provided accurate PP rates in all but 8, or over 97%, of these cases. The Rozeboom-based method provided accurate rates for 196, or 68%, of the 288 cases. The Rozeboom-based method was least accurate when the number of predictors was small and the expected R^2 was large. The adaptation of Sawyer's method used in the supplemental analysis did not provide consistent PP rates at all levels of expected R^2; however, across the number of predictors within expected R^2 values, the method did indeed provide consistent PP rates (which differs from the rules-of-thumb which varied both across predictors and across R^2 levels). The Sawyer method provided higher PP rates for the larger expected R^2 values, but never fell below .70.

CONCLUSIONS

The primary concern of this study was whether one or more of the methods examined provides consistently accurate precision power rates. That is, does any method provide reliable precision power rates regardless of the number of predictors and the value of expected R^2? The answer is that, assuming the researcher can make a reasonable estimate of the population ρ^2, the precision power method provides the most consistent precision power rates.

If the researcher cannot make a reasonable estimate of ρ^2, however, then no method works well. In other words, effect size is just as critical when choosing sample sizes in multiple regression as it is in other statistical methods, because all methods are inadequate when expected R^2 deviates too far from ρ^2. The results from these studies make it clear that researchers who hope to develop an efficient prediction model using multiple regression must be concerned with

the size of their derivation samples, starting with an appropriate expected R^2. It may be worth noting that although Stevens (1992a) suggested an effect size of $\rho^2=.50$ as a reasonable guess for the social sciences when a better estimate is unavailable, Rozeboom (1981) believes that $\rho^2=.50$ may be an upper bound and Cohen (1988) offers $\rho^2=.26$ as a large effect size. Of course, the best choice of effect size is based on evidence from the research literature or from past research experience.

Several questions remain regarding the selection of sample sizes for regression studies. The most critical question concerns the choice of a priori precision power rate. It is useful to remember that "for both statistical and practical reasons, then, one wants to measure the smallest number of cases that has a decent chance of revealing a significant relationship if, indeed, one is there" (Tabachnick & Fidell, 1989, p. 129). Although the precision power method provided PP rates within the accuracy criterion more frequently than did any other method, it typically provided values larger than the preset rate, which in turn implies a small excess of subjects. Specifically, the precision power method yielded PP greater than expected (but still within the accuracy range) in 259, or 90%, of the 288 supplemental conditions tested. In contrast, the Rozeboom-based method only reached the stated PP level in 47, or 16%, of the 288 cases. From a conservative viewpoint, this provides additional rationale for choosing the precision power method over the Rozeboom-based method. However, from a practical perspective, more subjects are required using the precision power method. For example, with eight predictors at expected $R^2=.25$, the precision power method recommends 279 subjects while the Rozeboom-based method suggests 248, or 31 fewer, subjects. The resulting PP rate for the precision power method was .824 and the PP rate for the Rozeboom-based method was .803. Clearly, the

Rozeboom-based method is preferable in this case, unless a better predetermined precision power choice can be made for the precision power method.

Because the precision power method provides highly accurate results, it is reasonable to take its conservative nature into account and use a lower a priori PP rate. Further examination of the data, particularly the supplemental data, indicates that it is not unreasonable to relax the preset precision power rate as the expected R^2 value decreases and the number of predictors increases. Indeed, this corresponds to the situations in which the Rozeboom-based method provides adequate power rates and preferable (i.e. more practical) sample size recommendations as compared to the precision power method. Based on review of average Stein estimates and standard deviations of those estimates, PP rates as low as .70 appear to provide reasonable results in the more extreme of these situations. Conversely, PP rates of .80 may be inadequate at the other end of the spectrum (high expected R^2 and few predictors). For example, for 3 predictors at an expected R^2=.80, the precision power method for PP=.80 recommends 14 subjects and provides and average PP=.801, average R^2=.824, and average Stein estimate of .673; however, only 24 subjects are recommended for precision power of .90, which provides PP=.906, R^2=.812, and Stein=.739 (with a much smaller standard deviation, 0.104 vs. 0.179). Clearly, effect size impacts the selection of sample size in complex ways. Such results make it more obvious as to why some scholars have recommended sample sizes of 100, 200, and even 500, no matter how many predictors, and others have suggested subject-to-variable ratios as large as 40:1 (e.g. Kerlinger & Pedhazur, 1972; Nunnally, 1978; Pedhazur, 1982; Tabachnick & Fidell, 1989).

Interestingly, what occurs when using the supplemental Sawyer method (i.e. $k=1+.1R^2$) is essentially what is described above. At expected R^2=.90, Sawyer provides precision power

consistently in the .92 range. As expected R^2 decreases, the Sawyer PP rates decrease gradually (but remain consistent across predictors within R^2 levels) until they apparently reach a level of about .66 when expected R^2=.10. Comparison of Stein estimates and standard deviations between the precision power method and the Sawyer method show that, at small R^2 values (e.g. 6 predictors and expected R^2=.20), one may be willing to use 106 fewer subjects (287 for the precision power method at PP=.80 vs. 181 for the supplemental Sawyer method) with 10% less precision power (.81 vs .71) to get an expected Stein estimate which is only 1% less (.174 vs. .163) and has a comparable standard deviation (.044 vs. .058). Of course, the flexibility of the precision power method allows researchers simply to choose an a priori precision power of .70 rather than forcing them to use the Sawyer method.

Summary

The seriousness of concern about sample sizes and precision power in regression is not obvious -- after all, researchers have shrinkage and cross-validity formulas available to "correct" for inadequate sample sizes. However, there is a theoretical relationship that both $E(R^2)$ and $E(R_c^2)$ converge toward ρ^2 as sample size increases (Herzberg, 1969). Indeed, further analysis of the current data shows that, within population ρ^2 ranges, there is a strongly negative relationship (from r=-.83, p<.001, when ρ^2=.05 to r=-.86, p<.001, when ρ^2=.75) between sample R^2 and the Stein cross-validity coefficient, R_c^2. Indeed, a similar negative relationship exists between R^2 and adjusted R^2, or R_a^2 (from r=-.55, p<.001, when ρ^2=.25 to r=-.67, p<.001, when ρ^2=.10). Therefore, even though R_a^2 differed from the population ρ^2 by an average of only .005 (standard deviation of .009), a larger sample size will still provide a better R_a^2 estimate of ρ^2. The good

news, if any, is that adjusted R^2 differed from ρ^2 by more than .02 in only 76, or 5%, of 1440 cases; further, all 76 of those cases had sample sizes less than 27.

The theoretical convergence noted by Herzberg (1969) may be best explained by an example that illustrates the differences between choosing a sample based on .80 precision power versus .80 statistical power. With four predictors and an expected R^2=.50, the precision power formula requires a sample of 55 and gives an average PP=.82, average statistical power of .999, average R^2=.531, average R_a^2=.493, and an average Stein R_c^2=.442; however, Cohen's method requires only 16 subjects but provides only an average PP=.37, average statistical power of .68, average R^2=.610, average R_a^2=.468, and average R_c^2=.208. The prediction model produced using a sample size from the precision power formula will better estimate both ρ^2 (using R_a^2) and ρ_c^2 (using R_c^2), and will provide more stable regression weights. Therefore, this model will predict better in future samples because the efficiency of a prediction model depends not on the estimates of ρ^2 and ρ_c^2, but on the stability of the regression coefficients.

The research presented in this paper is important for the reasons mentioned at the outset. Specifically, sample sizes for multiple linear regression, particularly when used to develop prediction models, must be chosen so as to provide adequate power both for statistical significance and also for generalizability of the model. It is well-documented and unfortunate that many researchers do not heed this guideline, choosing instead to abide by the rule cited by Olejnik (1984): use as many subjects as you can get. Possibly more tragic are the cases where researchers have used a groundless rule-of-thumb to choose their sample sizes or have neglected

to report an appropriate shrunken R^2, these studies probably provide inaccurate conclusions regarding the topics under investigation.

For whatever reasons, empirical study into power for multiple regression has been lacking. Rules-of-thumb have existed for decades with little empirical or mathematical support. Indeed, both the current study and its predecessor (Brooks & Barcikowski, 1994) have found very limited value for rules-of-thumb in regression. Additionally, sample size methods offered by Park and Dudycha (1974), Cohen (1988), and Gatsonis and Sampson (1989) were each found lacking in some way. Sawyer's (1982) original method simplifies into a rule-of-thumb if the inflation factor is set as a constant, which renders it only as useful as the rules-of-thumb described above. Two adaptations of Sawyer's method, which attempted to provide the method with an effect size, provided mixed results. The only method which provided consistently accurate power for generalizability was the new precision power method. The preliminary study's Rozeboom-based method finished as the second best method, but with problems that make it less desirable especially for large expected R^2 and few predictors.

It is hoped that both the evidence presented and the simplicity of the method developed in the current study encourage researchers to consider more seriously the issues of power and sample size for regression studies. Although power in regression studies may have additional meaning than for other statistical designs, it is no less important. Researchers must recognize the potential danger of choosing an inappropriate effect size (either implicitly or explicitly) or ignoring effect size entirely. Further, no statistical analysis or correction (such as an adjusted R^2) can repair damage caused by an inadequate sample. Researchers must remember that a sample must

not only be large enough, but that it must also be random and appropriately representative of the population to which the research will generalize (Cooley & Lohnes, 1971; Miller & Kunce, 1973).

References

Afifi, A. A., & Clark, V. (1990). Computer-aided multivariate analysis (2nd ed.). New York, Van Nostrand Reinhold.

Box, G. E. P., & Muller, M. E. (1958). A note on generation of normal deviates, AMS, 28, 610-611.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Brogden, H. E. (1972). Some observations on two methods in psychology. Psychological Bulletin, 77, 431-437.

Brooks, G. P., & Barcikowski, R. S. (April, 1994). A new sample size formula for regression. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Browne, M. W. (1975). Predictive validity of a linear regression equation. British Journal of Mathematical and Statistical Psychology, 28, 79-87.

Cascio, W. F., Valenzi, E. R., & Silbey, V. (1978). Validation and statistical power: Implications for applied research. Journal of Applied Psychology, 63, 589-595.

Cattin, P. (1980a). Estimation of the predictive power of a regression model. Journal of Applied Psychology, 65, 407-414.

Cattin, P. (1980b). Note on the estimation of the squared cross-validated multiple correlation of a regression model. Psychological Bulletin, 87, 63-65.

Chambers, J. M. (1977). Computational methods for data analysis. New York: John Wiley & Sons.

Claudy, J. G. (1972). A comparison of five variable weighting procedures. Educational and Psychological Measurement, 32, 311-322.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69, 161-182.

Darlington, R. B. (1990). Regression and linear models. New York: McGraw-Hill.

Dixon, W. J. (1990). BMDP statistical software manual to accompany the 1990 software release (Vol. 1). Berkeley, CA: University of California.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. Applied Psychological Measurement, 3, 387-399.

Dunn, O. J., & Clark, V. A. (1974). Applied statistics: Analysis of variance and regression. New York: John Wiley & Sons.

Ezekiel, M. (1930). Methods of correlational analysis. New York: Wiley.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. Psychological Bulletin, 106, 516-524.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26, 499-510.

Halinski, R. S., & Feldt, L. S. (1970). The selection of variables in multiple regression analysis. Journal of Educational Measurement, 7, 151-157.

Halperin, S. (April, 1976). Design of Monte Carlo studies. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 121 850)

Harris, R. J. (1985). A primer of multivariate statistics (2nd ed.). Orlando, FL: Academic Press.

Harwell, M. R. (April, 1990). Summarizing Monte Carlo results in methodological research. Paper presented at the meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 319 775)

Herzberg, P. A. (1969). The parameters of cross-validation. Psychometrika Monograph Supplement, 34(2, Pt. 2).

Hinkle, D. E., & Oliver, J. D. (1983). How large should a sample be? A question with no simple answer? Or.... Educational and Psychological Measurement, 43, 1051-1060.

Huberty, C. J. (1994, April). A note on interpreting an R^2 value. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 40, 101-112.

International Mathematical and Statistical Library. (1985). Stat/PC Library. Houston, TX: Author.

Johnson, N. L., & Leone, F. C. (1977). Statistics and experimental design in engineering and the physical sciences. New York: John Wiley & Sons.

Karian, Z. A., & Dudewicz, E. J. (1991). Modern statistical systems, and GPSS simulation: The first course. New York: Computer Science Press.

Kennedy, E. (1988). Estimation of the squared cross-validity coefficient in the context of best subset regression. Applied Psychological Measurement, 12, 231-237.

Kennedy, W. J., Jr., & Gentle, J. E. (1980). Statistical computing. New York: Marcel Dekker.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart, and Winston.

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violations of multisample sphericity. Psychological Bulletin, 110, 162-170.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1987). Applied regression analysis and other multivariate methods (2nd ed.). Boston: PWS-Kent.

Knuth, D. E. (1981). The art of computer programming: Vol. 2. Seminumerical algorithms (2nd ed.). Reading, MA: Addison-Wesley.

Lord, F. M. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample (Research Bulletin No. 50-40). Princeton, NJ: Educational Testing Service.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McNemar, Q. (1962). Psychological statistics (3rd ed.). New York: John Wiley & Sons.

Miller, D. E., & Kunce, J. T. (1973). Prediction and statistical overkill revisited. Measurement and evaluation in guidance, 6, 157-163.

Montgomery, D. C., & Peck, E. A. (1992). Introduction to linear regression analysis (2nd ed.). New York: John Wiley & Sons.

Morgan, B. J. T. (1984). Elements of simulation. New York: Chapman and Hall.

Murphy, K. R. (1982, August). Cost-benefit considerations in choosing among cross-validation methods. Paper presented at the meeting of the American Psychological Association, Washington, D.C. (ERIC Document Reproduction Service No. ED 223 701)

Nash, J. C. (1990). Compact numerical methods for computers: Linear algebra and function minimisation (2nd ed.). New York: Adam Hilger.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied linear statistical models: Regression, analysis of variance, and experimental designs (3rd ed.). Homewood, IL: Irwin.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin et al. (Eds.), Contributions to probability and statistics (pp. 322-330). Palo Alto, CA: Stanford University.

Norusis, M. J. (1988). SPSS-X advanced statistics guide (2nd ed.). Chicago: SPSS.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

Park, C. N., & Dudycha, A. L. (1974). A cross-validation approach to sample size determination for regression models. Journal of the American Statistical Association, 69, 214-218.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart, & Winston.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). Numerical recipes in Pascal: The art of scientific computing. New York: Cambridge University.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). Numerical recipes in FORTRAN: The art of scientific computing (2nd ed.). New York: Cambridge University.

Ray, A. A. (1982). SAS user's guide: Statistics, 1982 edition. Cary, NC: SAS Institute.

Ripley, B. D. (1987). Stochastic simulation. New York: John Wiley & Sons.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. British Journal of Mathematical and Statistical Psychology, 45, 283-288.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlations: A clarification. Psychological Bulletin, 85, 1348-1351.

Rozeboom, W. W. (1981). The cross-validational accuracy of sample regressions. Journal of Educational Statistics, 6, 179-198.

Rubinstein, R. Y. (1981). Simulation and the Monte Carlo method. New York: John Wiley & Sons.

Sampson, A. R. (1974). A tale of two regressions. Journal of the American Statistical Association, 69, 682-689.

Sawyer, R. (1982). Sample size and the accuracy of predictions made from multiple regression equations. Journal of Educational Statistics, 7, 91-104.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 84, 751-758.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), Contributions to probability and statistics. Palo Alto, CA: Stanford University.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. (1992a). Applied multivariate statistics for the social sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. (1992b, October). What I have learned (up to this point) or ruminations on twenty years in the field. Paper presented at the meeting of the Midwestern Educational Research Association, Chicago, IL.

Tabachnick, B. G., & Fidell, L. S. (1989). Using multivariate statistics (2nd ed.). New York: HarperCollins.

Thorndike, R. M. (1978). Correlational procedures for research. New York: Gardner.

Uhl, N., & Eisenberg, T. (1970). Predicting shrinkage in the multiple correlation coefficient. Educational and Psychological Measurement, 30, 487-489.

Weisberg, S. (1985). Applied linear regression (2nd ed.). New York: John Wiley & Sons.

West, L. J. (1990). Distinguishing between statistical and practical significance. Delta Pi Epsilon Journal, 32(1), 1-4.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 2, 440-451.

Table 1

Examples of Cross-Validation and Shrinkage Formulas

Formula	Attributed To:
$R_a^2 = 1 - \frac{(N-1)(1-R^2)}{(N-p)}$	Wherry (1931)
$R_a^2 = 1 - \frac{(N-1)(1-R^2)}{(N-p-1)}$	Wherry (1931); Ezekiel (1930); McNemar (1962); Lord & Novick (1968); Ray (1982, p. 69) [SAS]
$R_a^2 = R^2 - \frac{p(1-R^2)}{(N-p-1)}$	Norusis (1988, p. 18) [SPSS]
$R_a^2 = R^2 - \frac{p(1-R^2)}{(N-p')}$	Dixon (1990, p. 365) [BMDP][1]
$R_a^2 = R^2 - \frac{(p-2)(1-R^2)}{(N-p-1)} - \frac{2(N-3)(1-R^2)}{(N-p-1)(N-p+1)}$	Olkin & Pratt (1958)
$R_c^2 = 1 - \frac{(N-1)(N+p+1)(1-R^2)}{(N-p-1)N}$	Nicholson (1960) Lord (1950)
$R_c^2 = 1 - \frac{(N-1)(N-2)(N+1)(1-R^2)}{(N-p-1)(N-p-2)N}$	Stein (1960) Darlington (1968)
$R_c^2 = 1 - \frac{(N+p)(1-R^2)}{(N-p)}$	Rozeboom (1978)
$R_c^2 = 1 - \frac{(N+p+1)(1-R^2)}{(N-p-1)}$	Uhl & Eisenberg (1970) Lord (1950)

Note: R_a^2 represents an estimate of ρ^2; R_c^2 is an estimate of ρ_c^2.

[1] p'=p+1 with an intercept, p'=p if the intercept=0.

Table 2

Rules-of-Thumb for Sample Size Selection

Rule	Author(s)
$N \geq 10p$	Miller & Kunce, 1973, p. 162 Halinski & Feldt, 1970, p. 157 (for prediction if $R \geq .50$) Neter, Wasserman, & Kutner, 1990, p. 467
$N \geq 15p$	Stevens, 1992, p. 125
$N \geq 20p$	Tabachnick & Fidell, 1989, p. 128 ($N \geq 100$ preferred) Halinski & Feldt, 1970, p. 157 (for identifying predictors)
$N \geq 30p$	Pedhazur & Schmelkin, 1990, p. 447
$N \geq 40p$	Nunnally, 1978 (inferred from text examples) Tabachnick & Fidell, 1989, p. 129 (for stepwise regression)
$N \geq 50 + p$	Harris, 1985, p. 64
$N \geq 10p + 50$	Thorndike, 1978, p. 184
$N > 100$	Kerlinger & Pedhazur, 1973, p. 442 (preferably N>200)
$N \geq \frac{(2K^2-1) + K^2p}{(K^2-1)}$	Sawyer, 1982, p. 95 (K is an inflation factor due to estimating regression coefficients)

Note: In the formulas for sample size above, N represents the suggested sample size and p represents the number of predictors (independent variables) used in the regression analysis.

Table 3

Sample Sizes Suggested by Each Method for Each Level of Expected R^2

Number of Predictors	Method	Sample Size for: $E(R^2)=.75$	$E(R^2)=.50$	$E(R^2)=.25$	$E(R^2)=.10$
2	Precision Power ($\epsilon=.2R^2$)	13	33	93	273
	Rozeboom-based ($\epsilon=.2R^2$)	9	22	62	182
	Precision Power ($\epsilon=.05$)	33	63	93	183
	Rozeboom-based ($\epsilon=.05$)	22	42	62	122
	Park & Dudycha (p=.95)	23	42	62	119
	Park & Dudycha (p=.90)	18	31	45	85
	Sawyer	13	18	33	78
	30:1	60	60	60	60
	50 + 8p	66	66	66	66
	15:1	30	30	30	30
	Cohen	6	13	38	115
	Gatsonis & Sampson	11	20	45	135
3	Precision Power ($\epsilon=.2R^2$)	17	44	124	364
	Rozeboom-based ($\epsilon=.2R^2$)	13	33	93	273
	Precision Power ($\epsilon=.05$)	44	84	124	244
	Rozeboom-based ($\epsilon=.05$)	33	63	93	183
	Park & Dudycha (p=.95)	35	64	91	174
	Park & Dudycha (p=.90)	28	50	71	133
	Sawyer	17	24	44	104
	30:1	90	90	90	90
	50 + 8p	74	74	74	74
	15:1	45	45	45	45
	Cohen	7	14	44	130
	Gatsonis & Sampson	13	23	51	151
4	Precision Power ($\epsilon=.2R^2$)	22	55	155	455
	Rozeboom-based ($\epsilon=.2R^2$)	17	44	124	364
	Precision Power ($\epsilon=.05$)	55	105	155	305
	Rozeboom-based ($\epsilon=.05$)	44	84	124	244
	Park & Dudycha (p=.95)	44	82	117	220
	Park & Dudycha (p=.90)	37	66	93	173
	Sawyer	22	30	55	130
	30:1	120	120	120	120
	50 + 8p	82	82	82	82
	15:1	60	60	60	60
	Cohen	8	16	48	144
	Gatsonis & Sampson	14	25	55	165

Table 3 (continued)

8	Precision Power (ϵ=.2R^2)	39	99	279	819
	Rozeboom-based (ϵ=.2R^2)	35	88	248	728
	Precision Power (ϵ=.05)	99	189	279	549
	Rozeboom-based (ϵ=.05)	88	168	248	488
	Park & Dudycha (p=.95)	78	144	202	373
	Park & Dudycha (p=.90)	68	124	171	311
	Sawyer	38	53	98	233
	30:1	240	240	240	240
	50 + 8p	114	114	114	114
	15:1	120	120	120	120
	Cohen	12	20	61	183
	Gatsonis & Sampson	19	32	69	205
15	Precision Power (ϵ=.2R^2)	69	176	496	1456
	Rozeboom-based (ϵ=.2R^2)	65	165	465	1365
	Precision Power (ϵ=.05)	176	336	496	976
	Rozeboom-based (ϵ=.05)	165	315	465	915
	Park & Dudycha (p=.95)	131	261	331	600
	Park & Dudycha (p=.90)	118	214	292	524
	Sawyer	67	93	173	413
	30:1	450	450	450	450
	50 + 8p	170	170	170	170
	15:1	225	225	225	225
	Cohen	19	26	78	235
	Gatsonis & Sampson	27	42	88	256